Game-Theoretic Foundations for Norms

Guido Boella
Dipartimento di Informatica
Università di Torino-Italy
E-mail: guido@di.unito.it

Leendert van der Torre
Computer Science and Communications
University of Luxembourg
E-mail: leendert@vandertorre.com

Abstract

In this paper we study game-theoretic foundations for norms. We assume that a norm is a mechanism to obtain desired multi-agent system behavior, and must therefore under normal or typical circumstances be fulfilled by a range of agent types, such as norm internalizing agents, respectful agents fulfilling norms if possible, and selfish agents obeying norms only due to the associated sanctions.

1 Introduction

The relation between game theory and norms has received some attention. E.g., in a widely discussed example of the so-called centipede game, there is a pile of thousand pennies, and two agents can in turn either take one or two pennies. If an agent takes one then the other agent takes turn, if it takes two then the game ends. A backward induction argument implies that it is rational only to take two at the first turn. Norms and trust have been discussed to analyze this behavior, see [6] for a discussion.

Our approach to study this relation it to use game-theory for the foundations of normative systems. In artificial social systems or normative multi-agent systems, a social law or norm is a mechanism to achieve desired system behavior. Since in an open system it cannot be assumed that agents obey the norm, there has to be a control system motivating agents to obey the norm, by monitoring and sanctioning behaviors. Moreover, the system should not sanction without reason, as for example Caligula or Nero did in the ancient Roman times, as the norms would loose their force to motivate agents. Various ways to motivate agents including norms have been studied in economics, for example using the game-theoretic machinery to study the rationality of norms.

The research question of this paper is how to give game-theoretic foundations to norms such as obligations, permissions and counts-as conditionals. Most of our study is focussed on obligations, and therefore on incentives like sanctions and rewards.

We first consider the so-called partially controlled multi-agent system (PCMAS) approach of Brafman and Tennenholtz [3], one of the classical game-theoretical studies of social laws in so-called artificial social systems developed by Tennenholtz and colleagues, because incentives like sanctions and rewards play a central role in this theory. So-called controllable agents – agents controlled by the system programmer – enforce social behavior by punishing and rewarding agents, and thus can be seen as representatives of the normative system. For example, consider an iterative prisoner dilemma. A controlled agent can be programmed such that it defects when it happens to encounter an agent which has defected in a previous round.

The PCMAS model thus distinguishes between two kinds of agent interaction in the game theory, namely between two normal (so-called uncontrollable) agents, and between a normal and a controllable agent. We show in this paper that this makes it a very useful model to give game-theoretic foundations to norms. Whereas classical game theory is only concerned with interaction among normal agents, it is the interaction among normal and controllable agents which we use in our game theoretic foundations.

The PCMAS approach not only clarifies the design of punishments, but it also illustrates the iterative and multi-agent character of social laws. However, there are also drawbacks of the model, such that it cannot be used to give a completely satisfactory game-theoretic foundation for norms. We would like to express that a norm can be used for various kinds of agents, such as norm internalizing agents, respectful agents that attempt to evade norm violations, and selfish agents that obey norms only due to the associated sanctions. Therefore, as classical game theory is too abstract to satisfactorily distinguish among agent types, we consider also cognitive agents and qualitative game theory.

The layout of this paper is as follows. First we give some informal requirements, and we discuss how the Brafman-Tennenholtz model of PCMAS satisfies these requirements. Then we introduce a qualitative game-theory based on a logic for mental attitudes and cognitive agent theory, which we use to give game-theoretic foundations of obligations and permissions.

2 Requirements

Before we start, we would like to recall the role of incentives in economics. Consider the economist Levitt [8, p.18-20], discussing an example of Gneezy and Rustichini [5].

> Imagine for a moment that you are the manager of a day-care center. You have a clearly stated policy that children are supposed to be picked up by 4 p.m. But very often parents are late. The result: at day's end, you have some anxious children and at least a teacher must wait around for the parents to arrive. What to do?

> A pair of economists who heard of this dilemma – it turned out to be a rather common one – offered a solution: fine the tardy parents. Why, after all, should the day-care center take care of these kids for free?
>
> The economists decided to test their solution by conducting a study of ten day-care centers in Haifa, Israel. The study lasted twenty weeks, but the fine was not introduced immediately. For the first four weeks, the economists simply kept track of the number of participants who came late; there were, on average, eight pickups per week per day-center. In the fifth week, the fine was enacted. It was announced that any parent arriving more than ten minutes late would pay $3 per child for each incident. The fee would be added to the parents' monthly bill, which was roughly $380.
>
> After the fine was enacted, the number of late pickups promptly went ... up. Before long there were twenty late pickups per week, more than double the original average. The incentive had plainly backfired.
>
> Economics is, at root, the study of incentives: how people get what they want, or need, especially when other people want or need the same thing. Economists love incentives. They love to dream them up and enact them, study them and tinker with them. The typical economist believes the world has not yet invented a problem that he cannot fix if given a free hand to design the proper incentive scheme. His solution may not always be pretty– but the original problem, rest assured, will be fixed. An incentive is a bullet, a lever, a key: an often tiny object with astonishing power to change a situation.
>
> ...
>
> There are three basic flavors of incentive: economic, social, and moral. Very often a single incentive scheme will include all three varieties. Think about the anti-smoking campaign of recent years. The addition of $3-per-pack "sin tax" is a strong economic incentive against buying cigarettes. The banning of cigarettes in restaurants and bars is a powerful social incentive. And when the U.S. government asserts that terrorists raise money by selling black-market cigarettes, that acts as a rather jarring moral incentive.

The daycare example illustrates that an analysis should not naively restrict itself to economic incentives. The example illustrates that economic theory considers normative reasoning. Since classical decision and game theory are the main tools to study incentives in economics, we suggest that they are useful tools to study the role of normative incentives too. Though this is not uncommon in economics, most formal approaches to normative reasoning are developed regardless of game theoretic considerations. In the remainder of this section we list requirements for such an analysis.

The first requirement is that norms influence the behavior of agents. However, they only have to do so under normal or typical circumstances. For example, if other agents are not obeying the norm, then we cannot expect an agent to do so. This norm acceptance has been studied by [4], and in a game-theoretic setting for social laws by [14].

The second requirement is that even if a norm is accepted in the sense that the other agents obey the norm, an agent should be able to violate the norms. A normative multi-agent system is a "set of agents [...] whose interactions can be regarded as norm-governed; the norms prescribe how the agents ideally should and should not behave. [...] Importantly, the norms allow for the possibility that actual behavior may at times deviate from the ideal, i.e., that violations of obligations, or of agents' rights, may occur" [7].

In other words, the norms of global policies must be represented as soft constraints, which are used in detective control systems where violations can be detected, instead of hard constraints restricted to preventative control systems in which violations are impossible. The typical example of the former is that you can enter a train without a ticket, but you may be checked and sanctioned, and an example of the latter is that you cannot enter a metro station without a ticket.

Moreover, detective control is the result of actions of agents and therefore subject to errors and influenceable by actions of other agents. Therefore, it may be the case that violations are not often enough detected, that law enforcement is lazy or can be bribed, there are conflicting obligations in the normative system, that agents are able to block the sanction, block the prosecution, update the normative system, etc. A game-theoretic analysis can be used to study these issues of fraud and deception.

As the daycare example illustrates, the third requirement is that norms should apply to a variety of agent types. We assume that a norm is a mechanism to obtain desired multi-agent system behavior, and must therefore under normal or typical circumstances be fulfilled for a range of agent types, such as norm internalizing agents, respectful agents that attempt to evade norm violations, and selfish agents that obey norms only due to the associated sanctions. To distinguish these cases, we distinguish between the decision to count behavior as a violation, and to sanction it.

Given possible conditions for a norm, the fourth requirement is that norms are as weak as possible, in the sense that the norms should not apply in cases where this is undesired, and that sanctions should not be too severe. The latter is motivated by a classical economic argument due to Beccaria, which says that if sanctions are too high, they can no longer be used in cases where agents already have violated a norm. Sanctions should be high enough to motivate selfish agents, but they should not be too high.

3 Requirements in PCMAS

Several game-theoretic studies on social laws have been made by Tennenholtz and colleagues, for example based on off-line design of social laws [12], the emergence of conventions [13], and the stability of social laws [14]. The approach of Brafman and Tennenholtz [3] distinguishes between controllable and uncontrollable agents, analogous to the distinction between controllable and uncontrollable events in discrete event systems.

3.1 PCMAS

Controllable agents are agents controlled by the system programmer to enforce social behavior by punishing and rewarding agents. The game-theoretic model is the most common model for representing emergent behavior in a population. A single game consists of the usual payoff matrix. For example, the prisoner's dilemma is a two person game where each agent can either cooperate or defect.

Definition 1 *A k-person game g is defined by a k-dimensional matrix M of size $n_1 \times \ldots \times n_k$, where n_m is the number of possible actions (or strategies) of the m'th agent. The entries of M are vectors of length k of real numbers, called pay-off vectors. A joint strategy in M is a tuple $(i_1, i_2, \ldots, i_k)$, where for each $i \leq j \leq k$, it is the case that $1 \leq i_j \leq n_j$.*

An iterative game consists of a sequence of single games.

Definition 2 *A n-k-g iterative game consists of a set of n agents and a given k person game g. The game is played repetitively an unbounded number of times. At each iteration, a random k-tuple of agents play an instance of the game, where the members of this k-tuple are selected with uniform distribution from the set of agents.*

Efficiency is a global criterion for judging the "goodness" of outcomes from the system's perspective, unlike single payoffs which describe a single agent's perspective.

Definition 3 *A joint strategy of a game g is called* efficient *if the sum of the players pay-offs is maximal.*

New in the Brafman-Tennenholtz model are the notions of punishment and reward w.r.t. some joint strategy s, measuring the gain (benefit) or loss (punishment) of an agent if we can somehow change the joint behavior of the agents from a chosen efficient solution s to s'.

Definition 4 *Let s be a fixed joint strategy for a given game g, with pay-off $p_i(s)$ for player i; in an instance of g in which a joint strategy s' was played, if $p_i(s) \geq p_i(s')$ we say that i's* punishment *w.r.t. s is $p_i(s) - p_i(s')$, and otherwise we say that its* benefit *w.r.t. s is $p_i(s') - p_i(s)$.*

Agents may need to be constrained to behave in a way that is locally sub-optimal such that the multi-agent system is as efficient as possible. Brafman and Tennenholtz call such a constraint a social law. Then they informally define controlled agents:

> "Agents not conforming to the social law are referred to as *malicious agents*. In order to prevent the temptation to exploit the social law, we introduce a number of *punishing agents*, designed by the initial designer, that will play 'irrationally' if they detect behavior not conforming to the social law, attempting to minimize the payoff of malicious agents. The knowledge that future participants have of the punishment policy would deter deviations and eliminate the need for carrying it out. Hence, the punishing behavior is used as a threat aimed at deterring other agents from violating the social law. This threat is (part of) the control strategy adopted by the controllable agents in order to influence the behavior of the uncontrollable agents. Notice that this control strategy relies on the structural assumption that uncontrollable agents are expected utility maximizers."

They consider the design of punishments, and show, for example, necessary and sufficient conditions for the existence of a punishing strategy.

3.2 PCMAS as game-theoretic foundations for norms

We believe that PCMAS can be used to give game-theoretic foundations to norms, though Brafman and Tennenholtz do not use or consider the terminology of normative systems or deontic logic. The model fulfills our two requirements by explaining several aspects of norms, such as the fact that they can be used iteratively, that sanctions are associated to it, and that they can be applied to various kinds of agents.

In particular, a useful property of the PCMAS model is that it uses the game-theoretic machinery to study not only interaction among normal agents, but also interaction among the controlled agents and the normal agents. Since the controlled agents are representatives of the normative system, this means that the game-theoretic machinery is used to study the interaction among the normative system and the agents.

However, the emphasis on modeling uncontrollable agents as utility maximizers implies that they only obey the norm because they are afraid of the sanction. Thus the model does not fulfill the third requirement because it seems to exclude the possibility that an agent obeys the norm simply due to its existence. In social theory, for

example, agents have been studied which internalize norms in the sense that they incorporate norms as their own goal, or respectful agents trying to obey the norms without internalizing them.

Maybe the game-theoretic machinery can be extended to take such social agents into account. For example, a norm internalizing agent may be defined as an uncontrollable agent which simply copies the utility function of a punishing agent, and a respectful agent which avoids sanctions even when the number of punishing agents is too low, for example by assuming the number of punishing agents is much higher than it is in reality. They may for example be ashamed to be caught while driving without a train ticket.

However, such a solution does not seem very satisfactory. For the norm internalizing agents, they not only obey the norm but they also start to act as policemen, which seems to go to far. Moreover, even when punishment is low or absent, a respectful agent may obey the norm (as in the daycare example). There seem to be several alternative ways to define respectful agents, but they seem to have their own drawbacks.

Moreover, there are also some more technical problems. For PCMAS to give game-theoretic foundations to norms, we first have to define the syntax of a norm. Typically norms are expressed as modal sentences expressing that p is obliged, $O(p)$ in a deontic logic, or p is permitted, $P(p)$. Since in the PCMAS setting we have actions or strategies only, we define $O_i(\alpha, p)$ for agent i is obliged to do action α, otherwise he is sanctioned with punishment p. Since a punishment p is defined as $p_i(s) - p_i(s')$, the first problem is how to define the chosen efficient solution s'. It is implicit in the condition that in the situation in which no norm is violated, no agent is punished (the Nero/Caligula example of the introduction).

Finally, whether an obligation $O_i(\alpha, p)$ holds in PCMAS or not cannot be seen from the game's definition, but only from the behavior of the controlled agents. In other words, it can only be derived from the design of punishments not explicit in the game theory.

3.3 Qualitative game-theoretic foundations

Since agent types have been studied in qualitative game theories, it seems an obvious step to replace the utility maximizers by an alternative agent model. In the remainder of this paper we consider goal based cognitive agents, as they have been studied in philosophical logic, artificial intelligence, social theory and multi-agent systems. Since qualitative game theories are more closely related to logic than their quantitative predecessors, this will also facilitate the bridge to deontic logic.

For norm internalizing and respectful agents, we can use a qualitative game theory as follows:

- By replacing the agents utility function by goals, norm internalizing agents can

adopt (some of) the goals of the normative system. Note that this does not imply that the norm internalizing agents behave like the uncontrollable agents, because they may have other goals in addition, or other abilities or powers.

- Even when agents do not adopt the goals of the normative system, they can still take them into account. Moreover, by introducing a violation predicate in the logical language, we can distinguish between behavior which counts as a violation, and behavior which is sanctioned.

Qualitative game theories also have other advantages, though we do not study them in this paper. For example, they can be based on a more sophisticated notion of action, for example regarding causality or the observability of actions. Finally, a qualitative approach can be used to give a more detailed analysis of norms. It is less clear how PCMAS can be used to decompose the notion of norm or social law into separate components, as we will see in the following section.

4 Qualitative games among cognitive agents

In Boella and Lesmo's game-theoretic approach to norms [2], a rational definition of sanction-based obligations is given using classical game theory by representing the normative system as an agent. They model the normative system as a set of controlled agents, as in the PCMAS model, but they do not necessarily assume that they are controlled by the system programmer. We use a model of cognitive agents that is able to distinguish among norm internalizing agents, respectful agents that attempt to evade norm violations, and selfish agents that obey norms only due to the associated sanctions.

We have to be brief on technical details, and refer the reader to other work for the details. The important issue here is to give the flavor of cognitive agent theory, where the maximization of expected utility is replaced by maximization of achieved goals.

4.1 Input/output logic for mental attitudes

Makinson and van der Torre [9] define the proof theory of input/output logic as follows.

Definition 5 *Let L be a propositional language, let the norms in G be pairs of L $\{(\alpha_1, \beta_1), \ldots, (\alpha_n, \beta_n)\}$, read as 'if input α_1, then output β_1', etc., and consider the following proof rules strengthening of the input (SI), conjunction for the output (AND), weakening of the output (WO), disjunction of the input (OR), and cumulative transitivity (CT) and Identity (Id) defined as follows:*

$$\frac{(\alpha, \gamma)}{(\alpha \wedge \beta, \gamma)} SI \qquad \frac{(\alpha, \beta), (\alpha, \gamma)}{(\alpha, \beta \wedge \gamma)} AND \qquad \frac{(\alpha, \beta \wedge \gamma)}{(\alpha, \beta)} WO$$

$$\frac{(\alpha,\gamma),(\beta,\gamma)}{(\alpha\vee\beta,\gamma)}OR \qquad \frac{(\alpha,\beta),(\alpha\wedge\beta,\gamma)}{(\alpha,\gamma)}CT \qquad \frac{}{(\alpha,\alpha)}Id$$

The following four output operators are defined as closure operators on the set G using the rules above.

out_1*: SI+AND+WO*	*(simple-minded output)*
out_2*: SI+AND+WO+OR*	*(basic output)*
out_3*: SI+AND+WO+CT*	*(reusable output)*
out_4*: SI+AND+WO+OR+CT*	*(basic reusable output)*

Moreover, the following four throughput operators are defined as closure operators on the set G.

out_i^+*: out_i+Id (throughput)*

We write $out(G)$ for any of these output operations, and $out^+(G)$ for any of these throughput operations.

Semantics of input/output logics have been given for $out(G)$ in a classical Tarskian style (a model is a pair of sets of propositional valuations, with additional constraints) and for $out(G,A)=\{x \mid a\subseteq A,(a,x)\in out(G)\}$ in a more operational style. Moreover, extensions of input/output logics have been developed for contrary-to-duty reasoning [10] and for reasoning about weak and various kinds of strong permission [11].

The following definition extends constraints with a priority relation among norms, to resolve conflicts. Moreover, it introduces undercutter rules $a\rightarrow b$, which mean that if input a, then the output does *not* contain x. They are used to model permissions as exceptions to obligations.

Definition 6 *Let L be a propositional language, let G and H be two sets of pairs of L, let $\geq$: $2^{G\cup H}\times 2^{G\cup H}$ be a transitive and reflexive relation on subsets of these pairs, and let out be an output operation.*

- *A pair $\langle G',H'\rangle$ is consistent in a if $out(G',a)$ is consistent, and for each $(b,x)\in H'$, if $b\in out(G',a)$ then $x\notin out(G',a)$.*
- *$maxfamily(G,H,\geq,a)$ is the set of pairs $\langle G'\subseteq G,H'\subseteq H\rangle$ that:*
  1. *are consistent in a, and*
  2. *if $\langle G''\subseteq G,H''\subseteq H\rangle$ is consistent in a, $G'\subseteq G''$, and $H'\subseteq H''$, then $G''=G'$ and $H''=H'$;*

  In other words, it is maximal with respect to set inclusion among the consistent pairs;

- *$preffamily(G, H, \geq, a)$ is the set of pairs $\langle G', H' \rangle$ that:*
  1. *are in $maxfamily(G, H, \geq, a)$, and*
  2. *if $\langle G'' \subseteq G, H'' \subseteq H \rangle$ is in $maxfamily(G, H, \geq, a)$ and $G'' \cup H'' \geq G' \cup H'$, then $G'' = G'$ and $H'' = H'$;*

  In other words, it is maximal with respect to $\geq$;
- *$out_{\cap}(G, H, \geq, a) = \cap out(preffamily(G, H, \geq, a), \geq, a)$; In other words, using only the rules which occur in all elements of $preffamily$.*

4.2 Beliefs, goals, decisions, decision rule

To represent that agents are autonomous decision makers, we associate a set of decision variables with each agent. Decisions or actions are based on controllability from control theory or discrete event systems (not to be confused with controllable agents!). Moreover, each agent has four sets of rules, besides beliefs and goals also undercutters for beliefs and goals. Finally, each agent has a priority relation among these rules.

Definition 7 *Let L be a propositional logic based on the set of propositions X. A multi-agent system is a tuple $\langle A, B, G, C, H, AD, MD \geq \rangle$ where:*

- *the agents A, beliefs, B, goals G, belief undercutters C, and goal undercutters H are five disjoint sets;*
- *the agent description $AD : A \rightarrow 2^{B \cup G \cup C \cup H \cup X}$ is a function from agents to its beliefs, goals, undercutters, and decision variables;*
- *the mental description $MD : B \cup C \cup G \cup H \rightarrow L \times L$ is a function from the mental attitudes to input/output rules;*
- *the priority relation $\geq$: $A \rightarrow (2^{B \cup C} \times 2^{B \cup C}) \cup (2^{G \cup H} \times 2^{G \cup H})$ is a binary relation on sets of beliefs and goals.*

The qualitative decision rule is based on maximizing achieved goals, or minimizing unachieved goals.

Definition 8 *Given a multi-agent system, and let out^+ be a throughput operation for beliefs, and out an output operation for goals.*

- *A decision profile $\delta : (\cup_{a \in A} AD(a) \cap X) \rightarrow \{0, 1\}$ is a function from the decision variables to truth values; We represent δ by a logical formula;*
- *The expected effects of decision profile δ for agent a, $E(\delta, a)$, are $out^+_{\cap}(AD(a) \cap B, AD(a) \cap C, \geq, \delta)$;*

- *The unachieved goals according to agent a,* $U(\delta, a)$*, are* $\cap(G \setminus \{G' | \langle G', H' \rangle \in preffamily(G, H, \geq, E(\delta, a))\})$
- δ *is preferred to* δ' *iff* $U(\delta, a) \subset U(\delta', a)$

4.3 Agent types

The qualitative game theory works analogous to the classical game theory, where the maximization of expected utility is replaced by a minimization of unachieved goals. This more detailed model allows us to distinguish among various agent types. In this paper, we consider three agent types.

First, we consider norm internalizing agents. These uncontrollable agents incorporate some of the goals of the controllable agents. They thus behave like controllable agents, if it is in their power.

Second, respectful agents try to fulfill obligations when they can do so. We model this by making the violation conditions explicit in the controllable agents. They do not sanction directly, but they first determine whether observed behavior is a violation, and then associate a sanction with a violation. Respectful agents obey the norm, if they can, regardless of the sanction.

Third, selfish agents care only about the sanctions imposed by the controllable agents. They behave as traditional agents in economic theory.

5 Six clauses for obligation

For obligation and prohibition, we need at least six clauses. The first clause ensures that "respectful" agents internalizing the goals of the normative system will fulfill their obligation under typical circumstances, the second and third clause do so for "respectful" agents not internalizing the norm, and the other clauses do so for "selfish" types of agents. The first clause says that the obligation is in the desires and in the goals of a normative system b ("your wish is my command"). The second and third clause can be read as "the absence of x is considered as a violation". The association of obligations with violations is inspired by Anderson's reduction of deontic logic to alethic modal logic [1]. The third clause says that the normative system desires that there are no violations. The fourth and fifth clause relate violations to sanctions and assume that normative system b is motivated to apply sanctions only as long as there is a violation; otherwise the norm would have no effect. Finally, for the same reason, we assume in the last clause that the agent does not like the sanction.

Definition 9 (Obligation) *Let MAS be a multi-agent system, and* $G_a = AD(a) \cap G$*, etc. Agent* $a \in A$ *is obliged in MAS to decide to do* x *with sanction* s *if* Y *by controllable agent b, written as* $MAS \models O_{a,b}(x, s \mid Y)$*, if and only if:*

1. $Y \rightarrow x \in out(G_b)$*: if controllable agent* b *believes* Y*, then it desires and has as a goal that* x*.*

2. $Y \wedge x \rightarrow V_a(\neg x) \in out(G_b)$*: if controllable* b *believes* Y *and* $\neg x$*, then it has the goal* $V_a(\neg x)$*: to recognize* $\neg x$ *as a violation by agent* a*.*

3. $\top \rightarrow \neg V_a(\neg x) \in G_b$*: controllable agent* b *desires that there are no violations.*

4. $Y \wedge V_a(\neg x)\} \rightarrow s \in out(G_b$*: if controllable agent* b *believes* Y *and decides* $V_a(\neg x)$*, then it desires that it sanctions agent* a *with* s*.*

5. $Y \wedge \neg s \in out(G_b)$*: if controllable agent* b *believes* Y*, then it desires not to sanction,* $\neg s$*. The controllable agent only sanctions in case of violation.*

6. $Y \rightarrow \neg s \in out(G_a)$*: if agent* a *believes* Y*, then it desires* $\neg s$*, which expresses that it does not like to be sanctioned.*

6 Two clauses for permission

We do not define permissions as the absence of obligation, so-called negative permission, but as exceptions to obligations, a kind of positive permission. For a discussion on the issues involved in modeling permission, see [11]. Permission is simpler than obligation, since permissions cannot lead to violations and sanctions.

Here we distinguish between permission and entitlement or right. It is only due to entitlement that knowledge providers may be sanctioned when they do not permit a user to access documents, but the user itself cannot be a violator and be sanctioned due to its permissions to access a document. which distinguishes between users that are only permitted to access knowledge, and users that are also entitled to it in the sense that knowledge providers are obliged to permit them access. Games can be played to show that the clauses of permission are necessary, again for norm internalizing agents and other types of agents respectively.

Definition 10 (Permission) *Let MAS be a multi-agent system. Agent* $a \in A$ *is permitted to decide to do* x *if* Y *in* MAS *by controllable agent* b*, written as* $MAS \models P_{a,b}(x \mid Y)$*, if and only if:*

1. $Y \rightarrow x \in out(H_b)$*: if controllable agent* b *believes* Y*, then it does not have a goal that* x*.*

2. $Y \cup \{x\} \rightarrow \neg V_a(x) \in out(G_b)$*: if controllable agent* b *believes* Y *and* x*, then it does not want to count* x *as a violation.*

7 Summary

We show how the distinction between controllable and uncontrollable agents can be used to give game-theoretic foundations for norms. First, we discuss how the PCMAS model can be used to give such foundations for obligations. The main drawback of this model is that it is not clear how to distinguish among agent types. Another drawback is that it does not explain how to deal with other kinds of norms, such as permissions.

Second we discuss a straightforward extension of the PCMAS model using goal based reasoners instead of utility maximizers. We show also how game-theoretic considerations can be used to model permissions, based on a discussion by Bulygin. The use of qualitative game theory makes a bridge to deontic logic, which formalized logical relations among obligations and permissions.

A third kind of norms are constitutive norms known as counts-as conditionals, which define institutional facts in a normative system, for example which pieces of paper count as money, and which relations count as marriages. The game theoretic foundations of this kind of norms is subject of further research.

References

[1] A.R. Anderson. The logic of norms. *Logic et analyse*, 2, 1958.

[2] G. Boella and L. Lesmo. A game theoretic approach to norms. *Cognitive Science Quarterly*, 2(3-4):492–512, 2002.

[3] Ronen I. Brafman and Moshe Tennenholtz. On partially controlled multi-agent systems. *J. Artif. Intell. Res. (JAIR)*, 4:477–507, 1996.

[4] R Conte, C Castelfranchi, and F Dignum. autnomous norm acceptance. 1998.

[5] Gneezy and Rustichini. A fine is a price. *Journal of Legal Studies*, XXIX(1):1–18, 2000.

[6] M. Hollis. *Trust within Reason*. Cambridge University Press, 1998.

[7] A. Jones and J. Carmo. Deontic logic and contrary-to-duties. In D. Gabbay, editor, *Handbook of Philosophical Logic*, pages 203–279. Kluwer, 2002.

[8] Levitt and Dubner. *Freakonomics*. William Morrow, New York, 2005.

[9] D. Makinson and L. van der Torre. Input-output logics. *Journal of Philosophical Logic*, 29:383–408, 2000.

[10] D. Makinson and L. van der Torre. Constraints for input-output logics. *Journal of Philosophical Logic*, 30(2):155–185, 2001.

[11] D. Makinson and L. van der Torre. Permissions from an input/output perspective. *Journal of Philosophical Logic*, 32 (4):391–416, 2003.

[12] Y. Shoham and M. Tennenholtz. On social laws for artificial agent societies: off-line design. *Artificial Intelligence*, 73 (1-2):231 – 252, 1995.

[13] Y. Shoham and M. Tennenholtz. On the emergence of social conventions: modeling, analysis, and simulations. *Artificial Intelligence*, 94 (1-2):139 – 166, 1997.

[14] M. Tennenholtz. On stable social laws and qualitative equilibria. *Artificial Intelligence*, 102 (1):1–20, 1998.